

MULTIGRID METHODS FOR TENSOR STRUCTURED MARKOV CHAINS WITH LOW RANK APPROXIMATION

MATTHIAS BOLTEN*, KARSTEN KAHL* AND SONJA SOKOLOVIĆ*

Abstract. Tensor structured Markov chains are part of stochastic models of many practical applications, e.g., in the description of complex production or telephone networks. The most interesting question in Markov chain models is the determination of the stationary distribution as a description of the long term behavior of the system. This involves the computation of the eigenvector corresponding to the dominant eigenvalue or equivalently the solution of a singular linear system of equations. Due to the tensor structure of the models the dimension of the operators grows rapidly and a direct solution without exploiting the tensor structure becomes infeasible. Algebraic multigrid methods have proven to be efficient when dealing with Markov chains without using tensor structure. In this work we present an approach to adapt the algebraic multigrid framework to the tensor frame, not only using the tensor structure in matrix-vector multiplications, but also tensor structured coarse-grid operators and tensor representations of the solution vector.

Key words. multigrid method, tensor truncation, Tensor Train, Markov chains, singular linear system

AMS subject classifications. 65F10, 65F50, 60J22, 65N55

1. Introduction. Consider the transition rate matrix A of an irreducible continuous-time Markov chain and the task to find a state vector x such that

$$Ax = 0. \quad (1.1)$$

where $x \geq 0$ component wise and the sum of the components is equal to one, i.e., x can be interpreted as a vector of probabilities.

The matrix $A \in \mathbb{R}^{n \times n}$ is singular with corresponding rank $n - 1$ and has column sum zero, i.e., $\mathbf{1}^T A = 0$ where $\mathbf{1} = [1, \dots, 1]^T$. The solution of (1.1) is called the steady state vector and its existence and uniqueness (up to a scalar factor) is proven in [1]. Many continuous-time Markov chains arising in practical applications are known for the so called state space explosion, see, e.g., [14], especially those where the generator matrix A has tensor structure

$$A = \sum_{t=1}^T \bigotimes_{j=1}^J E_j^t, \quad (1.2)$$

where the matrices $E_j^j \in \mathbb{R}^{n_j \times n_j}$ describe local transitions in each of the J submodels and the other matrices $E_j^t \in \mathbb{R}^{n_j \times n_j}$ the synchronized transitions between the submodels. The sum over the Kronecker products of the matrices gives us the generator matrix of the whole system; cf. [11, 14]. Models of this type appear, e.g., in queuing theory [16, 17, 30] and analysis of stochastic automata networks [35, 41]. Note that most of the $E_j^t, t \neq j$ are the identity matrix and that an increase of the number of synchronized transitions results in an increase of the number of summands in (1.1). In addition, the dimension of A grows exponentially in the number of submodels J . Even for moderately large J the storage and computational complexity of forming A explicitly is prohibitive. Thus, in order to be able to do any computations for such models, all operations with A have to make use of the compact format (1.2). Due to

*Fachbereich C, Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, 42097 Wuppertal, Germany, {bolten,kkahl,sokolovic}@math.uni-wuppertal.de

the fact that $x \in \mathbb{R}^{\prod_{j=1}^J n_j}$ it is again difficult to calculate the exact solution due to memory constraints. If the solution x of (1.1) can be approximated by a vector \tilde{x} of the form

$$\tilde{x} = \sum_{i=1}^R \bigotimes_{j=1}^J x_j^i \quad (1.3)$$

with small R , the memory requirement for x is reduced from $\prod_{j=1}^J n_j$ to $R \sum_{j=1}^J n_j$. In addition, using this format matrix-vector multiplications can now be calculated with even smaller computational cost. While this assumption is certainly fulfilled for the extreme case of non-interacting networks, see Theorem 2, it is unclear whether this can be expected to be true in all applications of interest. Numerical experiments shown in section 5 suggest that this is the case at least for the Markov chains considered here, c.f. Fig. 3.

The structure of \tilde{x} can be interpreted as an outer vector product, which is equivalent to a J -way tensor in the canonical format with tensor rank R ; cf. [31, 32]. The approximation quality of \tilde{x} depends on its tensor representation; cf. [28]. In this work we focus on the Tensor Train (TT) format [37, 39, 40] and its performance in a multigrid setting. Recent publications which deal with multigrid methods for Markov chains [2, 8, 13, 18, 19, 20] do not discuss the use of tensors formats in a multigrid context, but show that for simple structured Markov chains a multigrid ansatz often works efficiently. Recently a multigrid approach has been proposed in [21] to obtain a low rank approximation to speed up the ALS algorithm.

In the last years there have been two main directions in the development of multigrid or multilevel algorithms for Markov chains. On the one hand there are methods based on smoothed aggregation multigrid [18, 19, 20], on the other hand the bootstrap algebraic multigrid framework has also been investigated in the context of computing the stationary distribution [2, 8]. In addition in [13] multigrid methods for structured Markov chains like the ones considered in this paper are studied. They are based on aggregation of the submodels and do not maintain the structure of (1.2). Apart from this multigrid approach there are publications which apply to problems with generator matrices of the form (1.2) which have a so called product form solution that can be then computed efficiently [12]. Note that one can interpret our approximation (1.3) of (1.1) as a product form solution if the number of summands is equal to one. Therefore it is obvious to interpret our approach as a combination and generalization of these two established solution techniques. Another recent approach for approximating the stationary distribution of tensor structured Markov chains by a low rank tensor was presented in [33], based on similar techniques for eigenvalue computations from [34]. This approach does, however, not use multigrid techniques.

Tensor techniques have been used in multigrid methods in different settings. In [3] a tensorized representation of multigrid methods similar to the one presented here was used to construct robust multigrid methods for partial differential equations, including singular perturbations. Later a multigrid method to solve large scale Sylvester equations including a low rank representation of the current approximate solution was presented in [25]. A related approach for the Lyapunov equation has been analyzed in [45].

The remainder of this work is structured as follows. In sections 2 and 3 we give an overview of the basic principles of multigrid methods and tensor formats, respectively. Section 4 gives a detailed explanation on how the individual ingredients of a multigrid

method are computed in such a way that the compact format of (1.2) is kept. Section 5 includes a variety of numerical tests and a discussion of the efficiency of our multigrid approach for different choices of building blocks and parameters. Concluding remarks and topics for future research are provided in section 6.

2. Multigrid basics. The building blocks of multigrid methods are smoothing schemes, the computation of the set of coarse variables, transfer operators and coarse grid operators. In the following there will be only an overview about what these concepts are and how they work together in a multigrid ansatz. For a detailed treatment and motivation we refer the reader to [43]. A prototype of a V -cycle multigrid method is given in Algorithm 1.

Algorithm 1: Multigrid V -cycle	
1	$v_l = \text{MG}(b_l, v_l)$
2	if <i>coarsest grid is reached</i> then
3	solve coarse grid equation $A_l v_l = b_l$.
4	else
5	Perform ν_1 smoothing steps for $A_l v_l = b_l$ with initial guess v_l
6	Compute the residual $r_l = b_l - A_l v_l$
7	Restrict $b_{l+1} = Q_l r_l$
8	$v_{l+1} = 0$
9	$e_{l+1} = \text{MG}(b_{l+1}, v_{l+1})$
10	Interpolate $e_l = P_l e_{l+1}$
11	$v_l = v_l + e_l$
12	Perform ν_2 smoothing steps for $A_l v_l = b_l$ with initial guess v_l
13	end

The smoothing process typically consists of a few iterations of a simple iterative method like weighted Jacobi, Gauß-Seidel or a Krylov subspace method like Richardson or GMRES; cf. [26, 27, 43]. Afterwards the error e of the current iterate v is calculated by approximately solving the residual equation $Ae = r$. This is done by performing the computations on a problem of smaller size on a so called coarse grid. To do so, the residual r and the operator A are restricted to a smaller space. In case it is the coarsest level in the hierarchy the restricted system is solved exactly, otherwise the process of smoothing and restriction is repeated until a grid is reached which is small enough for direct computations. Once the coarsest system is solved, the calculated error is interpolated to the next finer grid and added to the current iterate. A final smoothing operation is applied and the process is repeated until the finest grid is reached.

For notational simplicity we use a two grid notation whenever applicable, i.e., all quantities related to the coarse grid have subscript c . In cases where it is necessary to distinguish more than two subsequent grids we use a numbering of the grids as implied by algorithm 1.

In order to define restriction and interpolation operators one has to specify coarse variables. There are many different approaches available to define them, e.g., geometric coarsening [4], compatible relaxation [5, 7], aggregation [9] and some others which can be found in [43]. Typically these approaches split the variables into two sets, a set of fine variables \mathcal{F} and a set of coarse variables \mathcal{C} , which are used in the definition of restriction and interpolation. Assuming that such a splitting is chosen, restriction

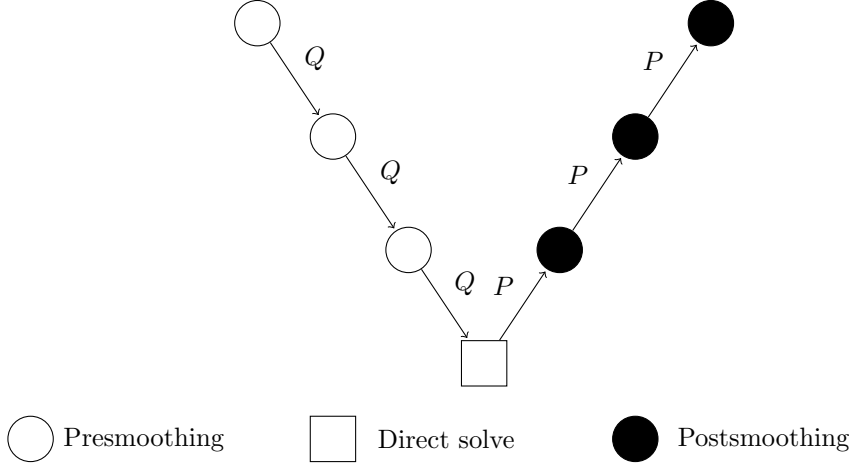


Fig. 1: Multigrid V-cycle: On each level, a presmoothing iteration is performed, then the problem is restricted to the next coarser grid. On the smallest grid, the problem is solved exactly by a direct solver. When interpolating back to the finer grids, postsmoothing iterations are applied on each level.

and interpolation operators are described by rectangular matrices. The restriction operator Q maps the residual to the coarse variables and the interpolation operator maps the calculated error correction from the coarse variables to the original space

$$Q : \mathbb{R}^{|\mathcal{C} \cup \mathcal{F}|} \rightarrow \mathbb{R}^{|\mathcal{C}|} \quad P : \mathbb{R}^{|\mathcal{C}|} \rightarrow \mathbb{R}^{|\mathcal{C} \cup \mathcal{F}|}.$$

There are also many different approaches to build these operators, e.g., least squares interpolation [6, 36], linear interpolation and others [43]. The coarse grid operator A_c is then formed by the Petrov–Galerkin construction $QAP \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{C}|}$ and the coarse residual is given by $r_c = Qr$. This is the standard choice in algebraic multigrid methods and gives us the approximated residual equation $A_c e_c = r_c$.

Instead of stopping on the second grid and solving $A_c e_c = r_c$ exactly, e.g., because A_c is still too large, one can again solve this system of equations by a two-grid approach. Iterating this idea ultimately yields a multigrid method, where only on the coarsest grid, i.e., the one with the smallest dimension, the corresponding system is solved exactly. This strategy is described in algorithm 1 and gives rise to a V-cycle depicted in figure 1. Other cycling strategies like W- or F-cycles [43] are also possible, but we do not consider them for the sake of simplicity. However, all ideas developed in this paper can also be used with these cycling strategies in a straight forward way.

3. Tensor basics. We define a tensor \mathcal{X} as follows.

DEFINITION 1. Consider the space $\mathbf{R} = \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_J}$ which is spanned by

$$\{v^{(1)} \otimes v^{(2)} \otimes \cdots \otimes v^{(J)} : v^{(j)} \in \mathbb{R}^{n_j}, 1 \leq j \leq J\}. \quad (3.1)$$

Then each element $\mathcal{X} \in \mathbf{R}$ is a J -way tensor and the generating products $v^{(1)} \otimes v^{(2)} \otimes \cdots \otimes v^{(J)}$ are called elementary tensors. Note, that one can interpret \mathcal{X} as a multidimensional array. We therefore label its entries by J indices as $\mathcal{X}(i_1, \dots, i_J)$.

The rank of a Tensor \mathcal{X} is the smallest integer r that satisfies

$$\mathcal{X} = \sum_{i=1}^r \mathcal{V}^{(i)} \quad (3.2)$$

for elementary tensors $\mathcal{V}^{(i)}$.

Note, that a one-way tensor is simply a vector and a two-way tensor is a matrix. Any elementary tensor has tensor rank 1.

We denote the mapping that maps a tensor \mathcal{X} to the corresponding vector by $\text{vec}(\mathcal{X})$. The ordering in which the entries appear in the vector is not crucial as long as it is consistent if there are multiple occurrences.

Closely related to tensors is the concept of the Kronecker product. As the Kronecker product of J vectors $v^{(1)}, v^{(2)}, \dots, v^{(J)}$, a vector of dimension $n_1 n_2 \dots n_J$, and the corresponding tensor $v^{(1)} \otimes v^{(2)} \otimes \dots \otimes v^{(J)}$ have the same entries we use the same symbol \otimes for tensor and Kronecker products. A useful property of the Kronecker product is given by

$$(A \otimes B)(C \otimes D) = AC \otimes BD. \quad (3.3)$$

when A, B, C, D are matrices such that all matrix-matrix products in (3.3) are defined.

The relation (3.3) is of particular interest in the situation that an elementary tensor $\mathcal{X} = x^{(1)} \otimes x^{(2)} \otimes \dots \otimes x^{(J)}$ and a Kronecker product of matrices $A^{(1)} \otimes \dots \otimes A^{(J)}$ denoted by A are given such that $A^{(j)} x^{(j)}$ is well-defined. In this situation we find for the multiplication $A\mathcal{X}$,

$$A\mathcal{X} = A^{(1)}x^{(1)} \otimes A^{(2)}x^{(2)} \otimes \dots \otimes A^{(J)}x^{(J)}. \quad (3.4)$$

Thus the product can again be described by an elementary tensor and it can be computed very efficiently by only computing matrix-vector products of small size n_j .

Motivated by the simple matrix-tensor multiplication for elementary tensors it is clear that it is beneficial to work with a tensor representation of \mathcal{X} instead of $\text{vec}(\mathcal{X})$, i.e., the full representation. There are many tensor formats (e.g., CP [15, 31, 32], H-Tucker [24, 29], Tensor Train [37, 39, 40]) with different approximation properties and matrix-tensor multiplications. Of these formats we focus on the Tensor Train (TT) format in the following and introduce its basic concepts next.

3.1. Tensor Train. A Tensor \mathcal{X} is in TT-format with TT-ranks r_0, \dots, r_J if each entry of \mathcal{X} is given as

$$\mathcal{X}(i_1, \dots, i_J) = G_1(i_1) \cdot G_2(i_2) \cdot \dots \cdot G_J(i_J) \quad (3.5)$$

with parameter dependent matrices $G_k(i_k) \in \mathbb{R}^{r_{k-1} \times r_k}$, $k = 1, \dots, J$. As $\mathcal{X}(i_1, \dots, i_J)$ is a scalar value, (3.5) implies that $r_0 = r_J = 1$. Note that the G_k can be interpreted as $r_{k-1} \times n_k \times r_k$ tensors and they are called cores of \mathcal{X} . The storage complexity of \mathcal{X} in format (3.5) is bounded by $(J-2)\hat{n}\hat{r}^2 + 2\hat{n}\hat{r}$, if $n_k \leq \hat{n}$ and $r_k \leq \hat{r}$ for $k = 1, \dots, J$.

To be able to use the TT-format in an iterative method, basic vector operations should perform cost efficiently. The addition of two tensors in the TT-format can be computed as follows: Given tensors \mathcal{X} and $\tilde{\mathcal{X}}$ with cores $G_k(i_k), \tilde{G}_k(i_k)$ respectively, then the cores of $\mathcal{X} + \tilde{\mathcal{X}}$ are given by

$$\begin{aligned} \hat{G}_1(i_1) &= \begin{pmatrix} G_1(i_1) & \tilde{G}_1(i_1) \end{pmatrix}, & \hat{G}_J(i_J) &= \begin{pmatrix} G_J(i_J) \\ \tilde{G}_J(i_J) \end{pmatrix}, \\ \hat{G}_k(i_k) &= \begin{pmatrix} G_k(i_k) & 0 \\ 0 & \tilde{G}_k(i_k) \end{pmatrix} \text{ for } k = 2, \dots, J-1. \end{aligned}$$

Thus, no arithmetic operations are necessary, but the TT-ranks (and therefore the storage complexity) of $\mathcal{X} + \tilde{\mathcal{X}}$ are higher than those of \mathcal{X} and $\tilde{\mathcal{X}}$. Therefore, a rounding (or truncation) procedure is necessary, which allows to approximate a given TT-tensor by a TT-tensor of lower rank. This truncation can be done in a numerically stable way with cost $\mathcal{O}(J\hat{n}\hat{r}^3)$ by using [37, Algorithm 2]. Another operation which is typically needed in iterative methods is the scaling of a vector (or tensor) by a scalar α . This can be done efficiently (and without changing the TT-ranks) by multiplying one of the cores by α . Another operation needed in our method is the Euclidean inner product. It can be evaluated by [37, Algorithm 4] with cost $\mathcal{O}(J\hat{n}\hat{r}^3)$.

A matrix A is in TT-format if each entry can be computed as

$$A(i_1, \dots, i_J; j_1, \dots, j_J) = M_1(i_1, j_1) \cdot M_2(i_2, j_2) \cdot \dots \cdot M_J(i_J, j_J) \quad (3.6)$$

with parameter dependent matrices $M_k(i_k, j_k) \in \mathbb{R}^{r_{k-1} \times r_k}$, $k = 1, \dots, J$. Note that a matrix A with TT-ranks $r_0 = r_1 = \dots = r_J = 1$ is just the Kronecker product of J matrices, i.e.,

$$A = M_1 \otimes M_2 \otimes \dots \otimes M_J. \quad (3.7)$$

Given a matrix A and a tensor \mathcal{X} in TT-format with all TT-ranks bounded by r , the matrix-vector product can be computed with arithmetic cost $\mathcal{O}(J\hat{n}^2\hat{r}^4)$ by using [37, Algorithm 5]. As the TT-ranks of A and \mathcal{X} multiply when performing this operation, it is crucial to use truncation afterwards.

In [37, Section 3] it is described how to efficiently convert a tensor in CP decomposition into the TT-format. Similar techniques can be used to convert a matrix A in representation (1.2) into the TT-format.

4. Method. Tensor-based methods are efficient only when the tensors appearing in the problem have low rank. If this is not the case, the tensor structure (1.2) of A can be used to implement fast matrix vector multiplication, but additional savings by using one of the described tensor formats for the iterate are not possible. Thus the success of our attempt to use a tensor format for the iterates depends crucially on the rank of the steady state vector. For non-interacting systems the stationary distribution has rank one, as stated by the following theorem, which is a basic result from linear algebra.

THEOREM 2. *For a non-interacting system, i.e., $E_j^t = I$ for $t \neq j$ and $T = J$ in (1.2) the stationary distribution p satisfies*

$$p = \bigotimes_{j=1}^J p^{(j)}, \quad (4.1)$$

where $p^{(j)}$ is the stationary distribution of E_j^j , i.e., it corresponds to a tensor of rank one.

The proof is simple, as the tensor product of the individual stationary distributions of the systems is the stationary distribution of the combined system.

Theorem 2 suggests that for systems with a low level of interaction the stationary distribution can be approximated by a tensor with low rank. Numerical experiments in, e.g., [33] as well as our experiments in section 5 show that in practice a rather low rank is often sufficient. Thus savings of both memory and compute time can be realized by using a truncated tensor format in an iterative method.

This result as well as the numerical experiments motivate the development of multigrid methods for system matrices A with tensor structure that allow for the efficient use of tensor formats as described in section 3. In the following we will specify the building blocks of a multigrid method according to the description in section 2. Our main goal in the following is to preserve the tensor structure of A in (1.2).

4.1. Smoother. Unlike in multigrid methods for matrices the tensor structure of the linear systems at hand rules out smoothers that require access to individual entries of the operator, e.g., Jacobi, which requires the diagonal, or Gauß-Seidel, which requires the lower triangular part. Unless there is a way to represent these parts of the operator in a tensor format as well, it is very difficult and prohibitively expensive to use these smoothers. Thus typical candidates that can be applied in a tensor environment are smoothers that only require multiplications of the iterate with the system matrix, scalar products, multiplication with scalars and addition of vectors (or their tensor representations). These requirements are fulfilled by, e.g., Richardson method or more generally any polynomial method and thus also by Krylov subspace methods. As the optimal smoothing parameter in the case of Richardson is hard to find we suggest to use a Krylov subspace method like GMRES with a fixed number of iterations as the prototype smoother in a tensor environment. As long as the number of iterations done in the smoother is small the overhead in storage and computation required for GMRES are negligible.

In case the system matrix allows for an explicit representation of the diagonal D or the upper and lower triangular matrices L and U with zero diagonal such that $A = D - L - U$, a Jacobi or Gauß-Seidel like smoother can be employed by approximately inverting $(D - L)$ using a method like AMEn [22, 23]. The effectiveness of such an approach hinges mainly on the conditioning of the linear systems that need to be solved in AMEn. Especially as it is known that the systems arising in AMEn for A become increasingly ill-conditioned for growing problem sizes, special care has to be paid when applying this approach.

4.2. Selection of coarse variables. The following proposition gives us a starting point for achieving our goal of maintaining the structure of A . For that purpose we assume that the restriction operator Q and the prolongation operator P itself possess the same tensor structure as A . The proposition can be proven using basic properties of the Kronecker product.

PROPOSITION 3. *Let A of the form (1.2) be given with $E_j^t \in \mathbb{R}^{n_j \times n_j}$. Let $P = \bigotimes_{j=1}^J P_j$ and $Q = \bigotimes_{j=1}^J Q_j$ with $P_j \in \mathbb{R}^{n_j \times n_j^c}$ and $Q_j \in \mathbb{R}^{n_j^c \times n_j}$ where $n_j^c \ll n_j$. Then the corresponding Petrov–Galerkin operator satisfies*

$$QAP = \sum_{t=1}^T \bigotimes_{j=1}^J Q_j E_j^t P_j. \quad (4.2)$$

A direct consequence of this proposition is that the set of coarse variables can be computed for the smaller problems described by the matrices E_j^* . However, it is not always straightforward how to do this. In case the matrices E_j^* originate from some geometric structure, like the overflow queuing network example in section 5.1, one can use an established coarsening algorithm. A matrix like in the Kanban system example in section 5.2 does not correspond to a connected graph, indeed there are several isolated states. In this case, coarsening can be done with respect to the

auxiliary matrices

$$\tilde{E}_j = \sum_{t=1}^T E_j^t, \quad (4.3)$$

which collect all transitions of the j -th subsystem and correspond to connected graphs again. Problems of the first kind can be thought of j individual Markov chains which are coupled in some way, but are also meaningful on their own, while the second kind are systems which have a high level of interaction where each individual subsystem is dysfunctional without interacting with the other systems. This will be seen in more detail in the description of the different test examples in section 5.

4.3. Transfer operators. Once the set of coarse level variables is selected, transfer operators P and Q have to be defined in accordance with our assumption on their structure. The way of forming P and Q as described in Proposition 3 directly implies that they are built with respect to the “local” operators E_j^* of smaller dimension and lifted to the full dimension of A via Kronecker products. The choice of the “local” interpolation and restriction operators depends on the considered problem: In case the problem is similar to the discretization of a partial differential equation, linear interpolation and a J -fold Kronecker product thereof is a reasonable choice. This is the case for the overflow queuing network problem considered in section 5 which bears similarity to an advection-diffusion problem. In principle all known constructions from algebraic multigrid can be used to construct the transfer operators, including adaptive techniques.

4.4. Coarse grid operator. The coarse grid operator is chosen as the Petrov–Galerkin operator, i.e., $A_c = QAP$, being the natural extension of the Galerkin operator that is the optimal coarse grid operator in the symmetric case.

Finally, on the coarsest grid the residual equation has to be solved. There are two ways to do this, either the coarsening results in a problem of a size where direct solving without exploiting the tensor structure is possible or by an iterative method (e.g., GMRES) that exploits the tensor structure. In both cases truncation has to be applied to obtain a low rank tensor representation of the solution.

In summary, by keeping in mind these considerations, we are able to maintain the tensor structure of the system matrix A throughout all levels of our multigrid method. This structure has two main advantages: First, only the small matrices need to be stored, substantially reducing the storage requirements and second an efficient matrix vector product can be performed if \mathcal{X} is given in a tensor format like the TT-format described in section 3.1 with low rank. While the chosen interpolation and restriction operators are of rank 1 the system matrix A usually is not. The first implies that multiplication with the grid transfer operators does not change the rank of a vector. The same holds for the coarse grid operator that is formed as the Petrov–Galerkin product. Though multiplication with the system matrix or the coarse grid operators usually results in a larger rank, as the resulting rank of a matrix vector product $A\mathcal{X}$, where both A and \mathcal{X} are in TT-format with ranks r_A and $r_{\mathcal{X}}$ respectively, is the product of the ranks $r_A r_{\mathcal{X}}$. Therefore, after each multiplication with an operator of rank larger than 1 truncation is employed.

5. Numerical Tests. In this section we illustrate how to choose the ingredients of a multigrid approach based on two examples. All computations were performed in MATLAB R2013a using the TT-Toolbox [38].

5.1. Overflow queuing network. The first example we consider is the so called Overflow Queuing Network. A finite number J of queues is given with a corresponding capacity k_i , $i = 1, \dots, J$. Customers arrive to an arbitrary queue q_i , $i = 1, \dots, J$ according to a Poisson process with rate λ_i , $i = 1, \dots, J$ and are served with an exponentially distributed service time with rate μ_i , $i = 1, \dots, J$. These local processes can be described by the following matrices

$$E_i^{(i,i)} = \begin{pmatrix} 0 & \mu_i & & & 0 \\ \lambda_i & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \mu_i \\ 0 & & & \lambda_i & 0 \end{pmatrix}, \quad i = 1, \dots, J. \quad (5.1)$$

Note that here and in the following, for a better understanding we denote the transitions by two-tuples instead of natural numbers. Here, the tuple (i, j) describes a transition which is initiated by queue i and affects queue j . The synchronized events appear if queue q_i has reached its capacity, i.e., q_i is full. In this situation an arriving customer has to enter q_{i+1} . If q_{i+1} is also full, the customer will enter the subsequent queue q_{i+2} and so on. If all of the subsequent queues are full, the customer will leave the system. For sake of simplicity we only describe the synchronized events in which the customer can enter the subsequent queue, because it is not full:

- if q_i is full, the customer leaves q_i , which leads to the following matrix

$$E_i^{(i,i+1)} = \begin{pmatrix} 0 & 0 & & & \\ 0 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 0 & 0 \\ & & & 0 & \lambda_i \end{pmatrix}, \quad i = 1, \dots, J-1.$$

- if q_i is full, q_{i+1} gets an additional customer, which leads to the following matrix

$$E_{i+1}^{(i,i+1)} = \begin{pmatrix} 0 & 0 & & & 0 \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 0 \\ 0 & & & 1 & 0 \end{pmatrix}, \quad i = 1, \dots, J-1.$$

All of the matrices E_i are of the dimension $n_i = k_i + 1$. For a detailed description of the model see [11]. Note that the choices of the parameters λ and μ affect whether more synchronized or local events appear. For a better understanding, consider 3 queues with capacity 32 and different choices of the parameters $\mu = [\mu_1, \dots, \mu_J]$ and $\lambda = [\lambda_1, \dots, \lambda_J]$, which can be seen in figure 2. The figures 2a, 2b and 2c show the distributions of the solutions for the corresponding parameter choice on each queue. In every figure we have three axes (one for each queue) and each axis describes the capacity. In case customers arrive slowly and are served rapidly (small λ , large μ) all queues are (almost) empty; cf. figure 2a. On the other hand, if customers

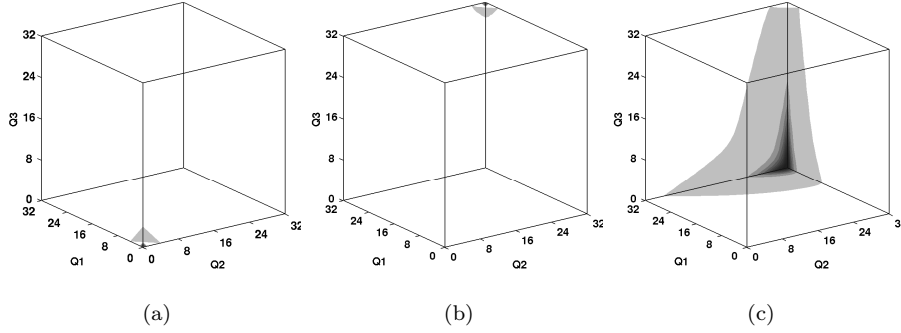


Fig. 2: Solution for parameters (a) $\mu = [1, 1, 1]$ and $\lambda = [0.1, 0.1, 0.1]$, (b) $\mu = [0.1, 0.1, 0.1]$ and $\lambda = [1, 1, 1]$, (c) $\mu = [0.25, 0.5, 1]$ and $\lambda = [0.5, 0.5, 0.5]$.

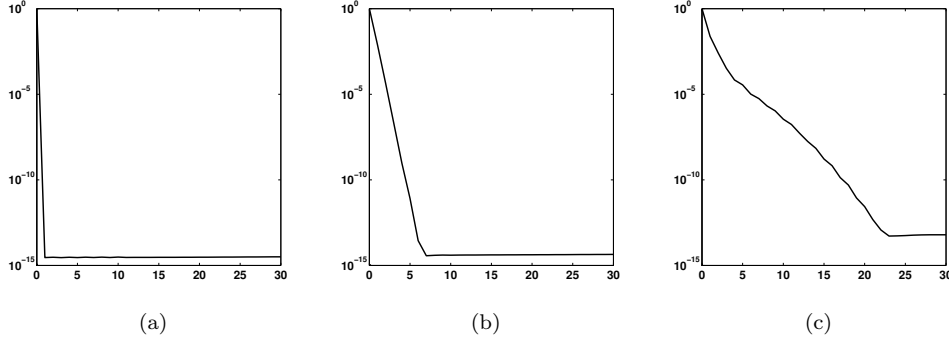


Fig. 3: Accuracy of rank R approximations (computed by truncation of the exact solution) for parameters (a) $\mu = [1, 1, 1]$ and $\lambda = [0.1, 0.1, 0.1]$, (b) $\mu = [0.1, 0.1, 0.1]$ and $\lambda = [1, 1, 1]$, (c) $\mu = [0.25, 0.5, 1]$ and $\lambda = [0.5, 0.5, 0.5]$.

arrive in short succession and are served slowly we obtain a steady state solution that corresponds to three full queues as can be seen in figure 2b. Both parameter sets have limited interaction, in the first situation no queue ever flows over which results in no interaction and in the second case almost every customer that is rejected at a queue is rejected at each subsequent queue as well, resulting effectively in no interaction as well. Only if the parameters are chosen such that slow and fast queues are mixed the steady state distribution becomes non-trivial as can be seen in 2c. In figure 3 we investigate fixed rank approximations of the solutions for the three different problems and observe that for the solution 2c a higher rank is needed than for the other ones due to the increased interaction of the queues. This observation confirms the assumption that a higher level of interaction in the model leads to a higher rank for its solution. Nevertheless, an accurate approximation of the solution is still obtainable with low rank.

One way to choose the ingredients for a multigrid approach is to take into account the underlying geometric structure and to preserve some important properties of the

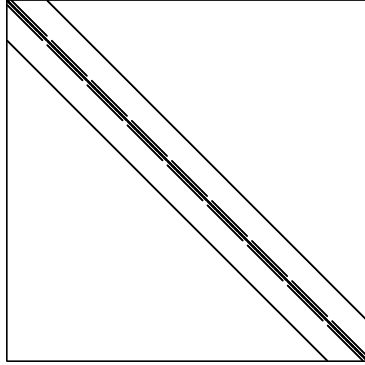


Fig. 4: Nonzero structure of overflow queuing network for $J = 3, k = [8, 8, 8]$.

problem matrix A on the coarse grids. In our case, the property that our matrix A has column sum zero should be kept. The structure of our overflow queuing problem for $J = 3, k_i = 8, i = 1, 2, 3$ is shown in Figure 4. Note that the nonzero structure is exactly the same as that of the standard finite difference discretization of the J -dimensional Laplace operator. A look at the geometric structure of the local problem in an overflow queuing problem (5.1) and the structural similarity to the Laplace operator motivated us to use full coarsening and direct interpolation for the interpolation operator, based on these local matrices, for each queue. To keep the column sum of the coarse grid matrix zero, we use linear restriction, because our coarse grid matrix is a Petrov–Galerkin operator and we then have $\mathbf{1}^T Q A P = \mathbf{1}^T A P = 0$ as the linear restriction operator Q fulfills $\mathbf{1}^T Q = \mathbf{1}$.

For the first numerical tests, we choose $J = 6, \lambda = [1.2, 1.1, 1.0, 0.9, 0.8, 0.7]$, and $\mu = [1, 1, 1, 1, 1, 1]$ with three different capacities $k = 8, 16$ and 32 . The dimension of each system is given by the formula $n = (k + 1)^J$. On every grid we use an implementation of GMRES that exploits the tensor structure in matrix vector products and inner products as the smoother and perform 3 pre- and postsmoothing steps for the first and second problem and 7 pre- and postsmoothing steps for the third problem. We use the Moore–Penrose pseudo inverse to solve the singular coarse grid system. As initial guess we use the smallest singular vector from the matrix corresponding to the smallest grid and interpolate it up to the original dimension, which can be done in the setup at low cost in a bootstrap fashion [6]. In experiments we could observe that this typically leads to a better starting point than a random initial guess. The availability of this starting guess is an additional advantage of a multigrid approach in this context. Note that after every V -cycle we normalize our iterate to guarantee that $\mathbf{1}^T x = 1$. We start with a maximal rank of 30 and increase the maximal rank by the factor $\sqrt{2}$ if after a V -cycle

- the iterate X_i has reached its maximal rank and
- the residual norm of X_i is not reduced or increased by 15% percent compared to the residual norm of the iterate X_{i-1} .

We are able to increase the rank adaptively, which is also an advantage of our method, because the rank of the state vector is not known a priori. The table 1 shows

J	k	n	levels	iter	max. rank	eff. rank	time
6	8	531,441	4	7	30	19.8	4.8
6	16	24,137,569	5	12	42	28.7	18.2
6	32	1,291,467,969	6	13	42	32.0	188.8

Table 1: Results of our multigrid algorithm for overflow queuing with 6 queues and varying capacity

the number of V -cycles, the effective and maximal rank and the time needed to reduce the residual norm $\|Ax\|$ below 10^{-7} for these three different problems. The effective rank of a TT-tensor \mathcal{X} is the number r_{eff} such that X requires the same amount of storage as a tensor with all TT-ranks equal to r_{eff} , see [42].

The numerical tests show that for problems with $J = 6$ the iteration number scales very nicely with the problem size. To guarantee that the used maximal rank in our tests is needed to approximate the solution and not only for the performance of our method, figure 5b shows the result of $\|AX\|$ where X is truncated which different truncation ranks. We see, that our method yields a solution with nearly "optimal" maximal rank. Figure 5a shows the convergence history of the three problems and the rectangles indicate in which iterations our method increases the maximal rank.

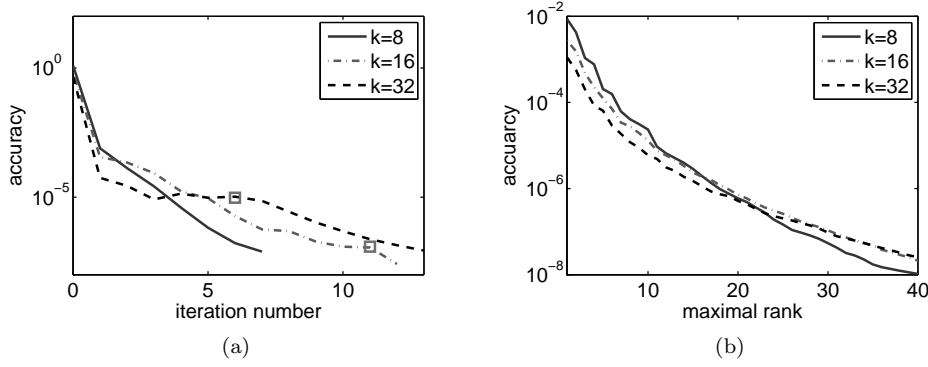
In order to compare our method with existing solution techniques we tested an implementation of alternating least squares with Tensor Train (ALS-TT) [33] for problems with queue sizes $J = 6$, $J = 7$, and $J = 8$. As ALS-TT requires an a priori rank-input we use the ranks determined by our multigrid method. The reduced problems which occur in ALS-TT were solved by the MATLAB backslash-solver. As a stopping criterion we chose the same accuracy requirement as for our multigrid method and aborted the iteration if the accuracy was not achieved after 4 hours. As can be seen in table 3 the absolute runtimes of ALS-TT are much worse than those of the multigrid algorithm, but due to the fact that this could be attributed to a non-optimal implementation of ALS-TT, we would rather like to point the attention to the scaling behaviour with respect to the queue size k of both methods. Doubling the queue size and thus increasing the system size by a factor of 64, the timings of the multigrid method grows by a factor of about 4. ALS-TT on the other hand requires more than 100 times as long. The slowing down of ALS-TT and other comparable methods like AMEn [33] with growing model size has been attributed in the literature to the ill-conditioning of the occurring linear systems that need to be solved in these methods. The fact that the multigrid method reduces the "local" systems to small sizes and typically also yields better conditioned coarse representations of these operators might explain the drastically improved scaling behaviour. The scaling of the two approaches with respect to the number of queues on the other hand is similar. Due to the fact that our coarsest system has the size $(n_i/2^{\#\text{levels}-1})^J$, i.e., it grows exponentially with J , we expect that using the pseudo-inverse to solve this system becomes infeasible for high dimensions J .

For the remaining test problems we do not report comparisons with ALS-TT as we expect it to behave similarly. However, note that Table 3 shows some other test results for higher dimensions, for which the behavior is analogous to the problems with $J = 6$.

5.2. Manufacturing system with Kanban control. The second example we consider is the so called Kanban model as it is discussed in [10]. A finite number J

J	k	n	multigrid time	ALS-TT time
6	8	531,441	4.8	87.9
6	16	24,137,569	18.2	> 4 hours
7	8	4,782,969	7.8	185.6
8	8	43,046,721	11.9	322.5

Table 2: Comparison of timings between multigrid algorithm and ALS-TT



of cells is given and each of them contains a machine, a bulletin board and an output hopper. Parts have to go through these machines in a certain order. We assume that this order is given by the numeration of the machines. Only a certain number $k_i, i \dots, J$ of parts can enter a machine, which is controlled by the so called Kanban tickets, i.e., k_i is the number of existing Kanban tickets for machine i . Each part which enters a machine gets a Kanban ticket and gives it back when it enters the subsequent machine. If no tickets are available in machine i the part has to wait in the output hopper of the previous cell $i - 1$ before entering the machine. The processing time and the time to move from one cell to the next are exponentially distributed with rate μ_i and rate ω_i , respectively. As in the previous example we want to distinguish the local events and the synchronized ones. Note that the synchronized events exist only between neighboring machines. For sake of simplicity we initially do not consider the first and the last machine. Each machine can be described by a Markov chain, where each state is characterized by three quantities:

- number of available tickets,
- number of parts being processed,
- number of parts waiting for the next machine.

In Figure 5 we illustrate the states of one machine with five tickets and characterize their transitions from one state to another as local or synchronized, where local transitions are the ones that are independent of the neighboring machines, i.e., the transition from the machine to the output hopper.

Ordering the states lexicographically and distinguishing the three different types of transitions, we find three different graphs corresponding to one machine, as depicted in Figure 6. The matrices describing the different transition types are exactly the transpose of the adjacency matrices of these graphs, with all nonzero entries equal to μ_i (for local transitions), ω_{i-1} (for transitions depending on the previous machine)

J	k	n	levels	iter	max. rank	eff. rank	time
7	8	4,782,969	4	9	30	20.7	7.8
7	16	410,338,673	5	15	30	21.2	27.6
8	8	43,046,721	4	10	30	21.8	11.9
8	16	6,975,757,441	5	17	30	22.1	40.1
9	8	387,420,489	4	10	30	22.76	16.0
9	16	118,587,876,497	5	20	40	26.1	66.7

Table 3: Results of our multigrid algorithm for overflow queuing with $\lambda = [1.2, 1.1, 1.0, 0.9, 0.8, 0.7, \dots]$, and $\mu = [1, 1, 1, 1, 1, 1, \dots]$

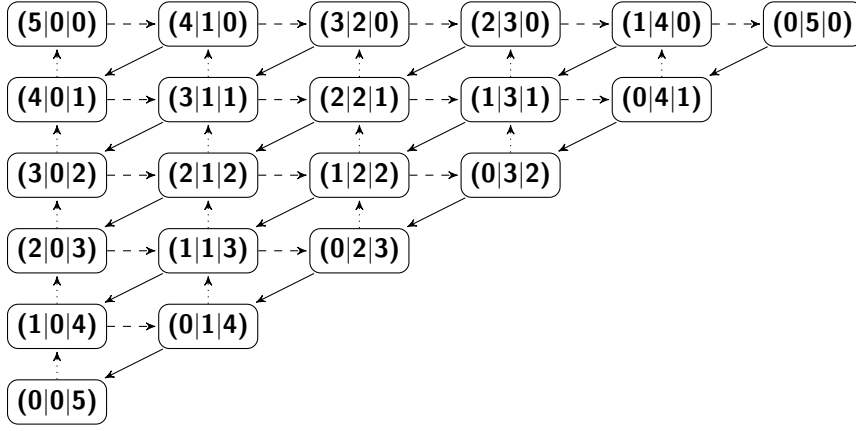


Fig. 5: Possible transitions between states of a machine with five Kanban tickets. Local transitions are depicted by solid arrows, synchronized transitions depending on the previous or subsequent machine by dashed or dotted arrows, respectively.

or ω_i (for transitions depending on the subsequent machine). In the local transition matrix, additionally the diagonal entries of nonzero columns are set to be $-\mu_i$. The corresponding matrices for a machine with two tickets are given as follows:

$$\begin{aligned}
 E_i^{(i,i)} &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\mu_i & 0 & 0 & 0 & 0 \\ 0 & \mu_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\mu_i & 0 & 0 \\ 0 & 0 & 0 & \mu_i & -\mu_i & 0 \\ 0 & 0 & 0 & 0 & \mu_i & 0 \end{pmatrix}, \quad E_i^{(i,i-1)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \omega_{i-1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \omega_{i-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & \omega_{i-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \\
 E_i^{(i,i+1)} &= \begin{pmatrix} 0 & 0 & \omega_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \omega_i & 0 \\ 0 & 0 & 0 & 0 & 0 & \omega_i \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad i = 2, \dots, J-1.
 \end{aligned}
 \tag{5.2}$$

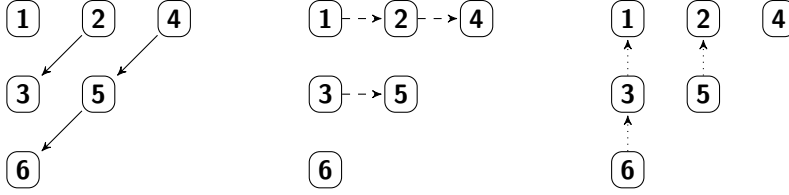


Fig. 6: Graphs corresponding to different transition types (left: local, middle: synchronized with previous machine, right: synchronized with subsequent machine) for a machine with two Kanban tickets. The states are ordered (and numbered) lexicographically, i.e., node 1 corresponds to $(2, 0, 0)$, node 2 corresponds to $(1, 1, 0)$ etc.

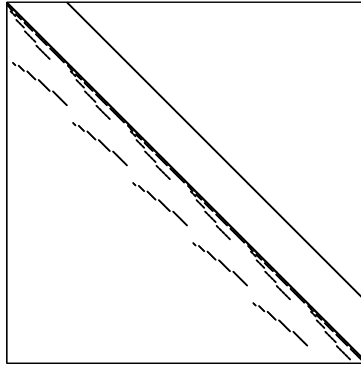


Fig. 7: Nonzero structure of Kanban model for $J = 3, k = [5, 5, 5]$.

The difference of the first and the last machine to the others is that they have only one neighbor, so the number of states of the corresponding Markov chain is smaller. To be more precise, the last machine lacks an output hopper (i.e., as soon as a part is finished its ticket is available), whereas the first machine has no bulletin boards (i.e., in case there are tickets available, a new part immediately enters the machine). Thus both the first and the last machine can be described by only two of the three state quantities. As such the states for the first and the last machine can also be read from figure 5. The first row of the state triangle describes the last machine and the first machine is given by the right-most diagonal.

The sparsity structure of the whole system with three machines and five tickets each is given in figure 7. In contrast to the overflow queuing model the graphs corresponding to the local matrices have some unreachable states, thus full coarsening with these matrices is not possible and in addition the geometric structure of these matrices does not bear a likeness to the structure of the system matrix depicted in figure 7, for example they have some zero columns. We thus base our coarsening approach on the accumulated local structure as shown in figure 5 and use a simple aggregation approach with constant basis function. On the left of figure 8 the selected aggregates for a problem with five tickets are shown. For the first and last machine the aggregates are chosen accordingly based on the identification with the right-most

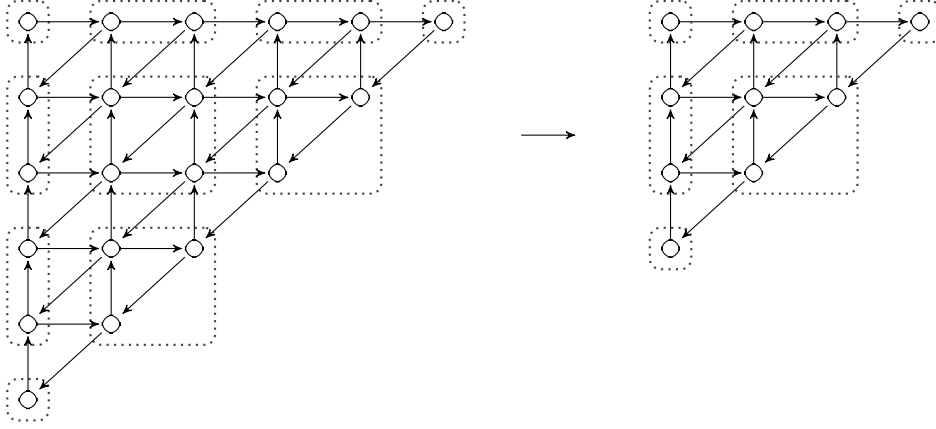


Fig. 8: Choice of aggregates in a three-level method for a machine with five tickets.

diagonal and first row, respectively. By enumerating the aggregates along the diagonals of the grid (starting in the upper left corner, just like the original numbering of the nodes in Figure 6) we preserve the connection structure of figure 5 and thus are able to continue aggregating according to this geometry to end up with a multilevel approach. On the right of figure 8 the choice of aggregates for further coarsening to a third level are shown. Note that the number of tickets in one machine should be $2^k + 1$ to allow for such an aggregation and we limit our analysis to such Kanban systems. This limitation is simply a technical issue and we suspect that irregular numbers of tickets can be dealt with by adapting the aggregation accordingly.

If the number of tickets is the same on every machine, which is denoted by k , the size of the system matrix is given by the formula $n = (k+1)^2 \cdot ((k+1)(k+2)/2)^{(J-2)}$. With the aggregation as described before we are able to reduce our problem to a size of $3^2 \cdot 6^{(J-2)}$. To reduce this to $2^2 \cdot 3^{(J-2)}$, we aggregate state 1 and 2, state 4 and 5 and state 3 and 6 (numbering as in figure 6) on the second to last level.

For our numerical tests we choose the aggregation approach as described before, therefore the choice of interpolation and restriction operator is given by standard aggregation interpolation (the restriction operator is just the transpose of the interpolation operator), see, e.g., [46], and the coarse grid matrix A_c is the corresponding (Petrov-)Galerkin operator as before. Note that the column sum of A_c is equal zero, because every column in the restriction matrix has only one entry with the value one. As a smoother, we use one approximate Gauß-Seidel iteration on each level, see [26, 43]. The lower and strictly triangular part L and U of our system matrix A is given in a Tensor representation, similar to A , see [44]. To apply the action of L^{-1} , we use the AMEn iteration from [22, 23] (with stopping accuracy 10^{-7} and a maximum of 20 sweeps) as implemented in the TT-Toolbox. Fortunately this turns out to be much easier than applying AMEn to the original problem. The reason may be that the matrix L is non-singular in contrast to A . Note that the linear systems that occur in AMEn for L are much better conditioned than the same systems for A . In addition, the structure of the Kanban control model, which is similar to a flow problem, also suggests that Gauß-Seidel is a good choice as a smoother, as it is known to be efficient for those problems.

J	k	n	levels	iter	max. rank	eff. rank	time
6	3	160,000	3	11	22	11.4	308.8
6	5	7,001,316	4	14	41	17.5	1136.5
6	9	915,062,500	5	23	59	29.6	12526.1
7	3	1,600,000	3	14	30	15.2	689.8
7	5	147,027,636	4	16	59	28.0	2051.1
8	3	16,000,000	3	16	42	21.9	1264.4
8	5	3,087,580,356	4	15	59	33.8	2783.8

Table 4: Results of our multigrid algorithm for Kanban model with varying number of machines and tickets and $\mu_i = 1, \omega_i = .1$ for all i .

The other ingredients of the method are the same as in example 5.1. Table 4 shows the results of some numerical tests for different numbers of machines and tickets. We can observe that the method works well for all tested problems. Again we see that a higher rank is needed with increasing dimensions of the problem. Apart from the rank increase we can again observe that the method scales reasonably well with growing problem size. The time the method takes is much higher than for the overflow test problems (mainly due to the use of the Gauß-Seidel smoother and the larger coarse grid systems) but this has to be expected as the Kanban problem is also much harder to solve for other solvers, see, e.g., [33].

6. Conclusion. We discussed how to develop multigrid methods for tensor-structured problems, more specifically for structured Markov chains, which on the one hand exploit the tensor structure for efficient matrix and vector operations and at the same time keep this structure intact across all grids. We illustrated the behavior of our method by investigating two standard model problems.

Topics for future research include exploring different coarsening strategies, e.g., aggregation across submodels (i.e., queues or machines for the overflow queuing or Kanban system, respectively) for large scale problems.

A further extension of this work is generalizing the structure-preserving multigrid approach to other classes of structured Markov chains, e.g., multi-server multi-queue models.

REFERENCES

- [1] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, 1994.
- [2] M. BOLTEN, A. BRANDT, J. BRANNICK, A. FROMMER, K. KAHL, AND I. LIVSHITS, *A bootstrap algebraic multilevel method for Markov chains*, SIAM J. Sci. Comput., 33 (2011), pp. 3425–3446.
- [3] S. BÖRM AND R. HIPTMAIR, *Analysis of tensor product multigrid*, Numer. Algorithms, 26 (2001), pp. 219–234.
- [4] A. BRANDT, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390.
- [5] ———, *General highly accurate algebraic coarsening*, Electron. Trans. Numer. Anal., 10 (2000), pp. 1–20.
- [6] A. BRANDT, J. BRANNICK, K. KAHL, AND I. LIVSHITS, *Bootstrap AMG*, SIAM J. Sci. Comput., 33 (2011), pp. 623–632.
- [7] J. BRANNICK AND R. FALGOUT, *Compatible relaxation and coarsening in algebraic multigrid*, SIAM J. Sci. Comput., 32 (2010), pp. 1393–1416.

- [8] J. BRANNICK, K. KAHL, AND S. SOKOLOVIC, *An adaptively constructed algebraic multigrid preconditioner for irreducible Markov chains*, submitted, preprint available as arXiv:1402.4005 (2014).
- [9] M. BREZINA, T. A. MANTEUFFEL, S. F. MCCORMICK, J. RUGE, AND G. SANDERS, *Towards adaptive smoothed aggregation (α SA) for nonsymmetric problems*, SIAM J. Sci. Comput., 32 (2010), pp. 14–39.
- [10] P. BUCHHOLZ, *Structured analysis approaches for large Markov chains*, Appl. Numer. Math., 31 (1999), pp. 375–404.
- [11] ———, *Multilevel solutions for structured Markov chains*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 342–357.
- [12] ———, *Product form approximations for communicating Markov processes*, in Quantitative Evaluation of Systems, 2008. QEST '08. Fifth International Conference on, sept. 2008, pp. 135–144.
- [13] P. BUCHHOLZ AND T. DAYAR, *Comparison of multilevel methods for Kronecker-based Markovian representations*, Computing, 73 (2004), pp. 349–371.
- [14] ———, *On the convergence of a class of multilevel methods for large sparse Markov chains*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 1025–1049.
- [15] J. D. CARROLL AND J. J. CHANG, *Analysis of individual differences in multidimensional scaling via an N -way generalization of Eckart-Young decomposition*, Psychometrika, 35 (1970), pp. 283–319.
- [16] R. CHAN, *Iterative methods for overflow queueing networks I*, Numer. Math., 51 (1987), pp. 143–180.
- [17] ———, *Iterative methods for overflow queueing networks II*, Numer. Math., 54 (1988), pp. 57–78.
- [18] H. DE STERCK, T. A. MANTEUFFEL, S. F. MCCORMICK, K. MILLER, J. RUGE, AND G. SANDERS, *Algebraic multigrid for Markov chains*, SIAM J. Sci. Comput., 32 (2010), pp. 544–562.
- [19] H. DE STERCK, T. A. MANTEUFFEL, S. F. MCCORMICK, Q. NGUYEN, AND J. RUGE, *Multilevel adaptive aggregation for Markov chains, with application to web ranking*, SIAM J. Sci. Comput., 30 (2008), pp. 2235–2262.
- [20] H. DE STERCK, T. A. MANTEUFFEL, S. F. MCCORMICK, J. RUGE, K. MILLER, J. PEARSON, AND G. SANDERS, *Smoothed aggregation multigrid for Markov chains*, SIAM J. Sci. Comput., 32 (2010), pp. 40–61.
- [21] H. DE STERCK AND K. MILLER, *An adaptive algebraic multigrid algorithm for low-rank canonical tensor decomposition*, SIAM J. Sci. Comput., 35 (2013), pp. B1–B24.
- [22] S. V. DOLGOV AND D. V. SAVOSTYANOV, *Alternating minimal energy methods for linear systems in higher dimensions. part I: SPD systems*, ArXiv e-prints, (2013).
- [23] ———, *Alternating minimal energy methods for linear systems in higher dimensions. part II: Faster algorithm and application to nonsymmetric systems*, ArXiv e-prints, (2013).
- [24] L. GRASEDYCK, *Hierarchical singular value decomposition of tensors*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2029–2054.
- [25] L. GRASEDYCK AND W. HACKBUSCH, *A multigrid method to solve large scale Sylvester equations*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 870–894.
- [26] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997.
- [27] W. HACKBUSCH, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, New York, NY, USA, 1994.
- [28] ———, *Tensor Spaces and Numerical Tensor Calculus*, Springer-Verlag, Berlin, Heidelberg, 2012.
- [29] W. HACKBUSCH AND S. KÜHN, *A new scheme for the tensor representation*, J. Fourier Anal. Appl., 15 (2009), pp. 706–722.
- [30] L. KAUFMAN, *Matrix methods for queuing problems*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 525–552.
- [31] H. A. L. KIERS, *Towards a standardized notation and terminology in multiway analysis*, J. Chemometrics, 14 (2000), pp. 105–122.
- [32] T. G. KOLDA AND B. W. BADER, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.
- [33] D. KRESSNER AND F. MACEDO, *Low-rank tensor methods for communicating Markov processes*, in Quantitative Evaluation of Systems, Gethin Norman and William Sanders, eds., vol. 8657 of Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 25–40.
- [34] D. KRESSNER, M. STEINLECHNER, AND A. USCHMAJEV, *Low-rank tensor methods with subspace correction for symmetric eigenvalue problems*, SIAM J. Sci. Comput., 36 (2014), pp. A2346–A2368.

- [35] A. N. LANGVILLE AND W. J. STEWART, *The Kronecker product and stochastic automata networks*, J. Comput. Appl. Math., 167 (2004), pp. 429 – 447.
- [36] T. A. MANTEUFFEL, S. F. MCCORMICK, M. PARK, AND J. RUGE, *Operator-based interpolation for bootstrap algebraic multigrid*, Numer. Linear Algebra Appl., 17 (2010), pp. 519–537.
- [37] I. V. OSELEDETS, *Approximation of $2^d \times 2^d$ matrices using tensor decomposition*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2130–2145.
- [38] ———, *MATLAB TT-Toolbox Version 2.2*, 2011. Available at http://spring.inm.ras.ru/osel/?page_id=24.
- [39] ———, *Tensor-Train decomposition*, SIAM J. Sci. Comput., 33 (2011), pp. 2295–2317.
- [40] I. V. OSELEDETS AND S. V. DOLGOV, *Solution of linear systems and matrix inversion in the TT-format*, SIAM J. Sci. Comput., 34 (2012), pp. A2718–A2739.
- [41] B. PLATEAU AND W. J. STEWART, *Stochastic automata networks*, in Computational Probability, Kluwer Academic Press, 1997, pp. 113–152.
- [42] D. SAVOSTYANOV, *QTT-rank-one vectors with QTT-rank-one and full-rank Fourier images*, Linear Algebra Appl., 436 (2012), pp. 3215 – 3224.
- [43] U. TROTTEBERG, C. OSTERLEE, AND A. SCHÜLLER, *Multigrid*, Academic Press, 2001.
- [44] E. UYSAL AND T. DAYAR, *Iterative methods based on splittings for stochastic automata networks*, Eur. J. Oper. Res., 110 (1998), pp. 166–186.
- [45] B. VANDEREYCKEN AND S. VANDEWALLE, *Local fourier analysis for tensorproduct multigrid*, AIP Conf. Proc., 1168 (2009), pp. 354–356.
- [46] P. VANĚK, *Fast multigrid solver*, Appl. Math., 40 (1995), pp. 1–20.